

Comparison of Multi-agent and Single-agent Inverse Learning on a Simulated Soccer Example

Xiaomin Lin¹ and Peter A. Beling² and Randy Cogill³

Abstract. We compare the performance of Inverse Reinforcement Learning (IRL) with the relative new model of Multi-agent Inverse Reinforcement Learning (MIRL). Before comparing the methods, we extend a published Bayesian IRL approach that is only applicable to the case where the reward is only state dependent to a general one capable of tackling the case where the reward depends on both state and action. Comparison between IRL and MIRL is made in the context of an abstract soccer game, using both a game model in which the reward depends only on state and one in which it depends on both state and action. Results suggest that the IRL approach performs much worse than the MIRL approach. We speculate that the underperformance of IRL is because it fails to capture equilibrium information in the manner possible in MIRL.

1 INTRODUCTION

The Multi-agent Reinforcement Learning (MRL) problem was first proposed by proposed by Littman [8] to address the limiting assumption in Reinforcement Learning (RL) that potentially responsive agents in a system area part of a passive environment. In a RL model, an agent can fully control the state transition process by taking actions on its own (though some stochastic variation is allowed); in a MRL problem, by contrast, the state transition process is determined by joint actions of all interacting rational agents. This essential difference complicates the MRL problems. As pointed out by Hu and Wellman [3], the concept of optimality, which is explicitly defined in IRL problems, loses its meaning in MRL problems since any agent's payoff depends on others' choices of action. In the absence of optimality, one can adopt as a solution concept the *Nash equilibrium*, in which each agent's choice is the best response to other agents' choices [3]. In fact, so far there is no agreement on a solution concept for a general MRL problem.

The first attempt to solve a MRL problem, given by Littman [8], made use of a Markov or stochastic game [11], which is an extension of game theory to *Markov Decision Process* (MDP)-like environments. However, only the special case of *two-player zero-sum games*, in which one agent's gain is always the other's loss, is considered in [8]. Hu and Wellman [3] extended Littman's work, proposing a *two-player general-sum* stochastic game framework for the MRL problem. Later MRL work has focused on the development of solution concepts and methods. For example, in [1] a weak condition where an agent can neither observe other agents' actions or rewards, nor knows the underlying game or the corresponding Nash equilibrium a priori is considered and a new MRL algorithm called the Weighted

Policy Learner (WPL) is proposed. Multi-agent learning in complex large distributed systems is also touched in [4], where it is noted that, although sophisticated multi-agent learning algorithms generally do not scale, it is possible to find restricted classes of games where simple efficient algorithms converge. Solution concepts for distributed, multi-agent planning problems that involve coordination games under weak information exchange models have been considered in [12, 16].

Inverse Reinforcement Learning (IRL), as the inverse learning problem for RL, has been studied extensively [10, 13, 6, 5, 14]. IRL aims to recover the reward function of the agent, in order to reason its behavior that is observed. Similarly, the inverse learning problems for MRL, termed MIRL, includes the problem of estimating the game payoffs being played, given only observations of the actions taken by the players. The reward function of an agent of interest recovered from an IRL approach, is in fact the mathematical expectation of all reward functions of other adaptive agents, which can be recovered from a MIRL approach. For example, if there are n adaptive agents in the environment and the one of interest is the k th of them, its reward at state s is

$$r^k(s, a^k) = \iint_{A^k} r(s, A^k) P(A^k|s) dA^k,$$

where

$$A^k = a^1 \dots a^{k-1} a^{k+1} \dots a^n$$

$$P(A^k|s) = p(a^1|s) \dots p(a^{k-1}|s) p(a^{k+1}|s) \dots p(a^n|s).$$

The above equation is valid in a general situation where the action space is continuous.

There exist several solution approaches for MIRL. Natarajan [9] presents an inverse reinforcement learning approach for multiple agents. However, that approach neither deals with competing agents nor considers a game-theoretic model. In [15], a form of the inverse equilibrium problem is discussed. However, that paper considers simultaneous one-stage games, rather than the sequential stochastic games. We have recently developed a Bayesian formulation for two-person zero-sum MIRL problems, in which an abstract soccer game, as a numerical example, is solved [2, 7].

Recall that MRL was proposed because the state transition dynamics remains unknown if other adaptive agent's actions are not taken into account, so that IRL is difficult to implement. However, in the two-person zero-sum MIRL model presented in [2], it is assumed that the two agents' policies of actions over all states, are known or observed. Therefore, a complete state transition matrix can be obtained. Then a question is naturally raised: is a MDP-based IRL approach able to solve the rewards recovery problem in a multi-agent

¹ University of Virginia, USA, email: xl5db@virginia.edu

² University of Virginia, USA, email: pb3a@virginia.edu

³ IBM Research, Ireland, email: rlc9s@virginia.edu

environment? This question is worth investigating deeply because if the answer is yes, there is no meaning to put addition effort into MIRL research.

The primary contribution of this work is to answer the above question. We first extend a Bayesian IRL approach, the idea of which was original proposed in [13], to infer the unknown rewards. Then we apply it on an abstract soccer game example, comparing the results with those obtained from the MIRL approach introduced in [2]. We consider two cases: one is that the unknown reward is only state dependent and the other We demonstrate that the results are much worse than that obtained from the MIRL approach introduced in [2]. This finding gives us an in-depth understanding the fundamentals of MIRL and substantiate the value of research on this topic.

The rest of the paper is structured as follows: Section 2 gives notation and concepts preliminary to MIRL, which is developed in the two-person zero-sum case in Section 3. Section 4 extends the Bayesian IRL approach proposed in [13] to a general form. Section 5 presents the soccer example and experiments in which the two methods are used to recover rewards. Section 6 provides an evaluation of the quality of the learned rewards in terms of game playing success in simulations of the soccer game. Finally, Section 7 offers concluding remarks.

2 PRELIMINARIES

A finite state two-person *discounted stochastic game* can be specified in terms of the state space $\mathcal{S} = \{1, 2, \dots, N\}$, the action spaces $\mathcal{A}_1 = \mathcal{A}_2 = \{1, 2, \dots, M\}$, reward functions $r^k : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \mapsto \mathbb{R}$ for each player $k \in \{1, 2\}$, transition probabilities $p(s'|s, a^1, a^2)$, and a discount factor $\gamma \in [0, 1)$.

In the MIRL model presented in [2], there are several rules governing the games played between two agents: First, each player has perfect knowledge of the other's rewards. Second, each player acts simultaneously in any state transition process and receives a reward depending on the starting state and their immediate actions. Third, they repeat the game over an infinite time horizon, aiming to accumulate maximum discounted rewards. In addition, all single-state games are zero-sum, which is, more specifically, $r^1(s, a^1, a^2) = -r^2(s, a^1, a^2)$. Due to the symmetry in rewards between the two players, we can simply use r to denote r^1 . Also, it is assumed that the *bipolicy* π of the two competing agents, which is a collection of both agents' policies of actions over all states, denoted (π^1, π^2) , is known.

The bipolicy-dependent, discounted expected sum of rewards of player 1 as a function of the initial state, which is known as the *value function*, can be formulated as:

$$V_\pi(s) = \sum_{t=0}^{\infty} \gamma^t E(\tilde{r}_\pi(s_t) | s_0 = s), \quad (1)$$

where s_t denotes the state of the game at stage t . $\tilde{r}_\pi(s_t)$ is the single-stage expected reward earned by player 1 at state s under bipolicy π , specifically,

$$\begin{aligned} \tilde{r}_\pi(s) &= \sum_{a^1, a^2} \pi^1(s, a^1) \pi^2(s, a^2) r(s, a^1, a^2) \\ &= [\pi^1(s)]^T r(s) \pi^2(s). \end{aligned} \quad (2)$$

Let r be the rewards vector of player 1, whose length is $M^2 N$. \tilde{r}_π can be expressed as

$$\tilde{r}_\pi = B_\pi r. \quad (3)$$

More details about B_π can be found in [7].

The state transition probability matrix under bipolicy π , G_π , is a $N \times N$ matrix with elements specified as

$$g_\pi(s'|s) = \sum_{a^1, a^2} \pi^1(s, a^1) \pi^2(s, a^2) p(s'|s, a^1, a^2). \quad (4)$$

A significant concept in two-person zero-sum MIRL is *minimax* bipolicy in which each player minimizes his own maximum loss. This is an equilibrium in that it has the property that neither player can change the game value in their favor given that the other player holds their policy fixed. In a two-person zero-sum stochastic game, we say that the two agents reach a minimax bipolicy if both of them employ a minimax strategy in every single-stage game.

3 BAYESIAN MIRL

In [2], the authors point out that the core of two-person zero-sum MIRL is the assumption that two agents reach a minimax bipolicy because both of them are rational. Using all terminologies and notations introduced in the preceding section, a convex quadratic program, in a Bayesian optimization setting, can be proposed for a general two-person zero-sum MIRL problem

$$\begin{aligned} \text{minimize: } & \frac{1}{2} (r - \mu_r)^T \Sigma_r^{-1} (r - \mu_r) \\ \text{subject to: } & (B_{\pi^2|a^1=i} - B_\pi) D_\pi r \leq 0 \\ & (B_{\pi^1|a^2=j} - B_\pi) D_\pi r \geq 0 \end{aligned} \quad (5)$$

for all $i \in \mathcal{A}_1$ and $j \in \mathcal{A}_2$, where μ_r is the mean of r and Σ_r is its covariance matrix. In the constraints of (5), $B_{\pi^k|a^{(3-k)}=i}$ ($k = 1, 2$) is conceptually similar to B_π , except that $B_{\pi^{(3-k)}|a^k=l}$ is constructed from a bipolicy in which player k always takes action l in any state while the other player still follows its original policy $\pi^{(3-k)}$. In addition, D_π can be expanded as

$$D_\pi = (I + \gamma P (I - \gamma G_\pi)^{-1} B_\pi), \quad (6)$$

where P is a $NM^2 \times N$ matrix with $p(s'|s, a^1, a^2)$ as its elements.

When it is known that r is only a function of state, we can use a simpler version of (5), as follows

$$\begin{aligned} \text{minimize: } & \frac{1}{2} (r - \mu_r)^T \Sigma_r^{-1} (r - \mu_r) \\ \text{subject to: } & (G_\pi - G_{\pi^2|a^1=i}) (I - \gamma G_\pi)^{-1} r \geq 0 \\ & (G_\pi - G_{\pi^1|a^2=j}) (I - \gamma G_\pi)^{-1} r \leq 0 \end{aligned} \quad (7)$$

for all $i \in \mathcal{A}_1$ and $j \in \mathcal{A}_2$, where $G_{\pi^{(3-k)}|a^k=l}$ ($k = 1, 2$) has a similar definition to that of G_π , except that $G_{\pi^{(3-k)}|a^k=l}$ is such a $N \times N$ state transition matrix that player k always takes action l in any state while the other player still follows its original policy $\pi^{(3-k)}$, the elements of which are, more specifically,

$$g_{\pi^{(3-k)}|a^k=l}(s, a^{(3-k)}) = \sum_{a^{(3-k)}} \pi^k(s, l) \pi^{(3-k)}(s, a^{(3-k)}) p(s'|s, l, a^{(3-k)}). \quad (8)$$

The theoretical validation of (5) and (7) are detailed in [2, 7].

4 BAYESIAN IRL

We will address the multi-agent inverse problem from the perspective of IRL. Qiao and Beling [13] propose a Bayesian optimization program based on the assumption that the agent's reward function is only

state dependent. In this section, we will extend this idea and formulate a more general program where the case that reward is state and action dependent is considered. Although we now turn to the MDP framework, most of the terminologies and notations introduced in Section 2 will still be adopted here, unless otherwise specified.

As stated before, we will focus on player 1's rewards. However, we are now tasked to recover $r_{\pi^2}(s, a^1)$, which is the expected value of $r(s, a^1, \pi^2(s))$ in case player 2 employs policy π^2 , specifically,

$$r_{\pi^2}(s, a^1) = \sum_{a^2} r(s, a^1, a^2) \pi^2(s, a^2). \quad (9)$$

For simplicity, we will just use r to denote the column vector whose element is $r_{\pi^2}(s, a^1)$, as follows:

$$r_{\pi^2} = \underbrace{(r_{\pi^2}(s_1, a_1^1), r_{\pi^2}(s_2, a_1^1), \dots, r_{\pi^2}(s_N, a_1^1))}_{r_{a_1^1}}, \dots, \underbrace{(r_{\pi^2}(s_1, a_M^1), r_{\pi^2}(s_2, a_M^1), \dots, r_{\pi^2}(s_N, a_M^1))}_{r_{a_M^1}})^T.$$

Note that the length of r here is MN , which is different from the one defined in Sections 2 and 3.

We define player 1's Q-function of state s and action a^1 under policy π^1 , $Q_{\pi^1}(s, a^1)$, to be the expected return from state s , taking action a^1 and thereafter following its original policy.

$$Q_{\pi^1}(s, a^1) = r_{\pi^2}(s, a^1) + \gamma \sum_{s'} p(s'|s, a^1) V_{\pi^1}(s'), \quad (10)$$

and its value function in state s is

$$\begin{aligned} V_{\pi^1}(s) &= \sum_{a^1} Q_{\pi^1}(s, a^1) \pi^1(s, a^1) \\ &= \tilde{r}_{\pi^1}(s) + \gamma \sum_{s'} g_{\pi}(s'|s) V_{\pi^1}(s'), \end{aligned} \quad (11)$$

where

$$\tilde{r}_{\pi^1}(s) = \sum_{a^1} r_{\pi^2}(s, a^1) \pi^1(s, a^1). \quad (12)$$

Hence (11) can be written in matrix notation as

$$V_{\pi^1} = \tilde{r}_{\pi^1} + \gamma G_{\pi} V_{\pi^1}, \quad (13)$$

where

$$\tilde{r}_{\pi^1} = C_{\pi^1} r, \quad (14)$$

and C_{π^1} is a $N \times NM$ matrix constructed from π^1 , whose i th row is,

$$\left[\underbrace{0, \dots, 0, \pi^1(i, 1), 0, \dots, 0}_N, \underbrace{\dots}_{(M-2)N}, \underbrace{0, \dots, 0, \pi^1(i, M), 0, \dots, 0}_N \right]$$

Thus

$$V_{\pi^1} = (I - \gamma G_{\pi})^{-1} C_{\pi^1} r. \quad (15)$$

The policy π^1 is optimal for player 1 in the sense that

$$V_{\pi^1}(s, \pi^1(s)) \geq Q_{\pi^1}(s, i), \quad (16)$$

for all $i \in \mathcal{A}_1$ and $s \in \mathcal{S}$, which means that in every state s , it is better (or equivalent) for player 1 to employ strategy $\pi^1(s)$ than

following any pure strategy. (16) can be expended as

$$\begin{aligned} \tilde{r}_{\pi^1}(s) + \gamma \sum_{s'} P_{s\pi^1(s)}(s'|s) V_{\pi^1}(s') \\ \geq r(s, i) + \gamma \sum_{s'} P_{sa^1=i}(s'|s) V_{\pi^1}(s'), \forall i \in \mathcal{A}_1. \end{aligned} \quad (17)$$

In (17), note that $P_{s\pi^1(s)}(s'|s) = g_{\pi}(s'|s)$ and $P_{sa^1=i}(s'|s) = g_{\pi^2|a^1=i}(s'|s)$. Expressing the above equation in matrix notation gives

$$\tilde{r}_{\pi^1} + \gamma G_{\pi} V_{\pi^1} \geq r_{a^1=i} + \gamma G_{\pi^2|a^1=i} V_{\pi^1} \quad (18)$$

where $r_{a^1=i} = C_{a^1=i} r$ and $C_{a^1=i}$ can be constructed from a pure policy where player 1 will take action i in any state. Substituting (15) and (14) into (18), gives

$$(F_{a^1=i}^{\pi^1} - C_{a^1=i}) r \geq 0, \quad (19)$$

where

$$F_{a^1=i}^{\pi^1} = [\gamma (G_{\pi} - G_{\pi^2|a^1=i}) (I - \gamma G_{\pi})^{-1} + I] C_{\pi^1}, \quad (20)$$

for all $i \in \mathcal{A}_1$.

To establish a Bayesian setting for IRL, we need to assign a prior distribution on the reward vector of player 1, $f(r)$. Let $p(\pi^1|r)$ denote the likelihood of observing player 1's policy π^1 when its true reward is r . We model $p(\pi^1|r)$ by

$$p(\pi^1|r) = \begin{cases} 1, & \text{if } Q_{\pi^1}(s, \pi^1(s)) \geq Q_{\pi^1}(s, i), \forall i \in \mathcal{A}_1 \\ 0, & \text{otherwise.} \end{cases} \quad (21)$$

The posterior distribution of the unknown rewards for a given observed policy π^1 is now

$$f(r|\pi^1) \propto p(\pi^1|r) f(r),$$

Hence

$$p(\pi^1|r) \propto \begin{cases} f(r), & \text{if } Q_{\pi^1}(s, \pi^1(s)) \geq Q_{\pi^1}(s, i), \forall i \in \mathcal{A}_1 \\ 0, & \text{otherwise.} \end{cases}$$

By assuming that r is Gaussian distributed, $r \sim \mathcal{N}(\mu_r, \Sigma_r)$, we can develop a standard optimization program with the posterior of r being the objective function and (19) being the constraint. Specifically,

$$\begin{aligned} \text{minimize: } & \frac{1}{2} (r - \mu_r)^T \Sigma_r^{-1} (r - \mu_r) \\ \text{subject to: } & (F_{a^1=i}^{\pi^1} - C_{a^1=i}) r \geq 0, \end{aligned} \quad (22)$$

for all $i \in \mathcal{A}_1$. In the above formulation, μ_r is the mean of the unknown reward vector as a prior, and Σ_r is its covariance matrix.

5 NUMERICAL EXPERIMENTS

In this section, we will apply the Bayesian IRL to the abstract soccer game introduced in [2, 7], and compare the results with those obtained from MIRL.

The game is played on a 4×5 grid as depicted in Figure 1. We use A and B to denote two players, and the circle in the figures to represent the ball. Each player can either stay unmoved or move to one of its neighborhood squares by taking one of 5 actions in each turn: N (north), S (south), E (east), W (west), and *stand*. Each player can only take one action in a single time period, and both of them act

simultaneously. If both players land on the same square in the same time period, the ball is exchanged between the two players with probability $\beta = 0.6$, which is known to the observer. There are in total 800 states in this model, corresponding to the positions of the players and ball possession. Each player aims to maximize its expected points scored, subject to a discount factor of $\gamma = 0.9$.

Both players attempt to dribble the ball into specific squares representing their opponent's goal. Player A attempts to score by reaching squares 6 or 11 with the ball, and player B attempts to score by reaching squares 10 or 15. Once a point is scored, the players take the positions shown in Figure 1 and ball possession is assigned randomly.

Obviously, the two rational players are playing a zero-sum stochastic game. As stated in Section 3, the bipolicy that they follow is a minimax bipolicy, and is known in this example. We are tasked to recover the reward structure of A, and thereafter infer which squares A must reach in order to score a point (the goal squares).

	1	2	3	4	5
	6	7	8	9	10
	11	12	13	14	15
	16	17	18	19	20

Figure 1. Soccer game: initial board

5.1 State Dependent Rewards

In the above game model, the reward is only state dependent. We will apply MIRL and IRL methods to recover A's rewards.

We use a Gaussian prior, where the mean reward assigns 0.8 point to player A in every state where A has possession of the ball and -0.8 point in every state where player B has possession of the ball. The covariance matrix of the prior is assumed to be an identity matrix, because without the knowledge of point structure, the correlation between different reward is not clear.

Results from these experiments are shown in Figure 2. In the figure, red circles represent the true reward, green triangles represent rewards learned from IRL, and blue stars represent rewards learned from MIRL. Examination of the figure shows a qualitative advantage for MIRL in that, in aggregate, the blue stars lie substantially closer to the red circles than do the green triangles. In Section 6 we assess the quality of learned rewards in terms of the quality of the forward policy that can be learned from them.

5.2 State and Actions Dependent Rewards

We now complicate the soccer game by allowing a "shoot" action. In addition to the available 5 actions, each player who has the ball can take a shot toward their opponent's goal at any position, with a *probability of succesful shot* (PSS) distribution over all positions in the field shown in Table 1. Assume that each one's PSS is independent of its opponent's position. One major difference from the

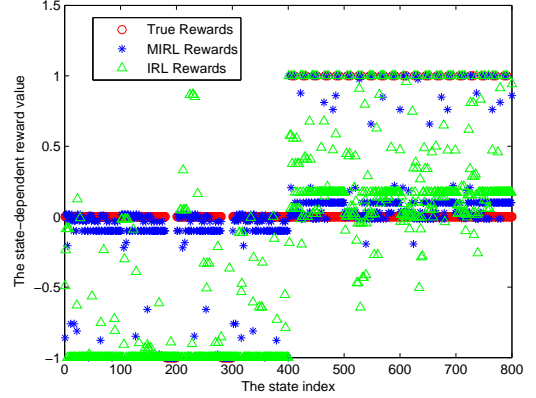


Figure 2. State-dependent rewards recovery results

simple model is that now the reward to be covered depends on both state and actions. Recall that IRL is used to infer $r(s, a^1)$ and MIRL will infer $r(s, a^1, a^2)$. We can compare the two results by calculate $r(s, a^1)$ from (9) in section 4. In both of the methods, we need to

	PSS = 1	PSS = 0.7	PSS = 0.5
A	6, 11	1, 7, 12, 16	2, 8, 13, 17
B	10, 15	5, 9, 14, 20	4, 8, 13, 19
	PSS = 0.3	PSS = 0.1	PSS = 0
A	3, 9, 14, 18	4, 10, 15, 19	5, 20
B	3, 7, 12, 18	2, 6, 11, 17	1, 16

Table 1. Original PSS distribution of each player

assign a mean and a covariance matrix for the prior. We can also develop three types of means based off of our knowledge of this game, as the following:

- *Weak Mean*: the same as the one described in section 5.1;
- *Median Mean*: guessing that A's goal might be among the rightmost squares, or squares 5, 10, 15 and 20, and symmetrically, B's goal might be among the leftmost squares, or squares 1, 6, 11 and 16, we assign 1 point to A whenever A has the ball and is in the four rightmost squares, and -1 point to A whenever B has the ball and is in four rightmost squares. Also, when A has the ball and takes a shot, no matter where she is, we assign 0.5 point to A. Otherwise, no points will be assigned to A.
- *Strong Mean*: we have a good guess of A's point distribution, except for its PSS distributions. So comparing to *median mean*, the only difference is that now the potential goal area includes only 2 squares (square 6 and 11 for A and square 10 and 15 for B), rather than 4 squares, for both players.

The covariance matrix of the reward vector encodes our belief of the structure of the prior. We can come up with a covariance matrix encapsulating some internal information subject to our knowledge of the relationship between rewards,

1. When A has the ball and takes a shot, the PSS depends only on A's position in the field.
2. In any state when A has the ball, the reward for A for any non-shoot action is a state-dependent constant.

We name this covariance matrix *Strong Covariance Matrix*, in order to distinguish it from the simple identity matrix we used in the simple game model.

Figures 3-5 show, for the various experiments, original rewards (red circles), rewards learned from IRL (green triangles), and rewards learned from MIRL (blue stars). It can be seen that in each figure there is a considerable overlap in distribution between the true rewards and MIRL rewards. The recovered rewards from IRL, by contrast, tend to lie far away from the true rewards.

We also check the recovery of A's PSS, present results in Figure 6-8. All these results show that compared to the MIRL approach, the IRL method is unable to give a reasonable recovery of A's PSS.

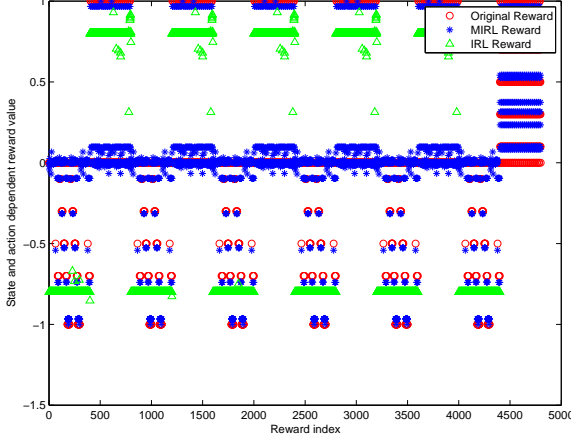


Figure 3. MIRL vs IRL on rewards: weak mean and strong covariance

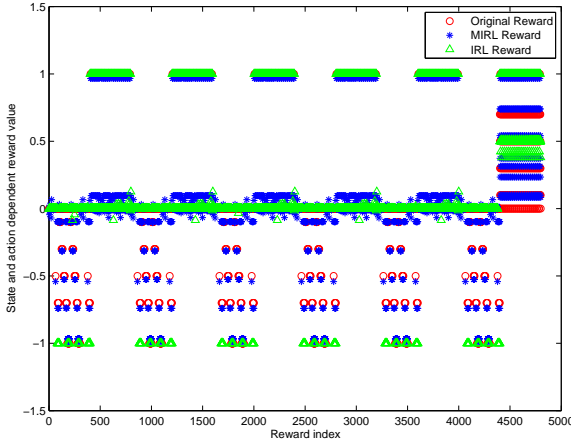


Figure 4. MIRL vs IRL on rewards: median mean and strong covariance

6 MONTE CARLO SIMULATION USING RECOVERED REWARDS

In this section, we measure the quality of learned rewards in terms of the quality of the forward solution that they induce. Let A employ the IRL rewards and B employ the MIRL rewards, and both believe that their own reward function is the true one. Being rational, both of them will employ a minimax bipolicy based off of their own rewards. Another criteria to evaluate the rewards quality is to apply them in a different environmental setting, e.g. $\beta = 1$. We will simulate games between A and B when $\beta = 0.6$ and $\beta = 1$, and compare the win-lose results of cases where different sets of rewards are employed.

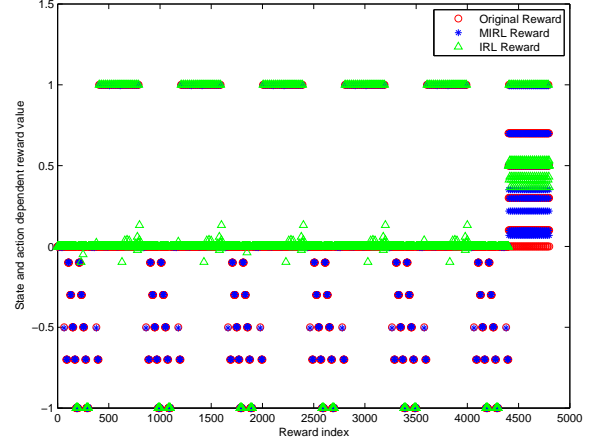


Figure 5. MIRL vs IRL on rewards: strong mean and strong covariance

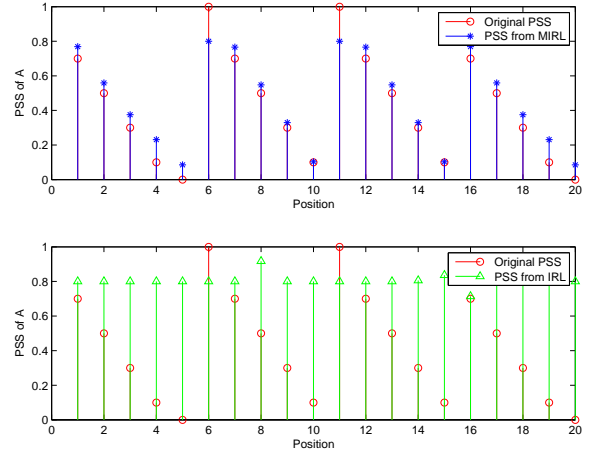


Figure 6. MIRL vs IRL on PSS: weak mean and strong covariance

The simulation results are presented in Table 2. In this table, the first column is the type of rewards that A and B employ to develop their minimax policies. A comparison only occurs when the two players use the same type of rewards on their own. *WM*, *MM*, *SM* and *SC* stand for *weak mean*, *median mean*, *strong mean*, and *strong covariance matrix*, respectively. The rest columns are the simulation results of 5000 rounds of games between A and B in cases where β being 0, 0.6 and 1. For a more clear comparison, we only count those game episodes ending in win-lose outcomes. Each column gives the winning percentage of B in each different rewards set they use. Our description of the game model indicates that A and B are supposed to be equal in match. However, this simulation results shows that B gets a big edge on A. We can conclude from that rewards learned from MIRL beat those learned from IRL in quality.

	% won ($\beta = 0$)	% won ($\beta = 0.6$)	% won ($\beta = 1$)
WM & SC	100	100	100
MM & SC	77.13	63.23	62.30
SM & SC	100	100	100

Table 2. A vs B games simulation results

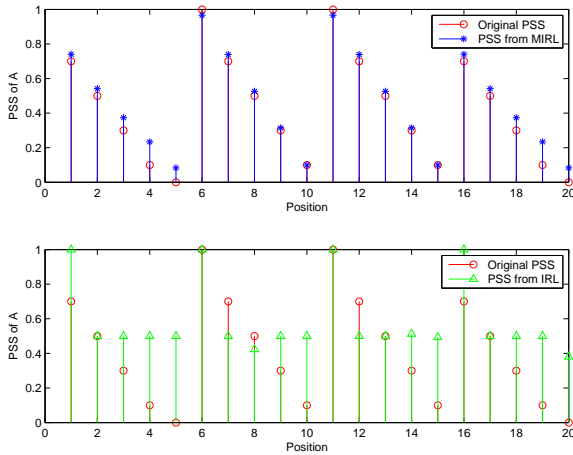


Figure 7. MRL vs IRL on PSS: median mean and strong covariance

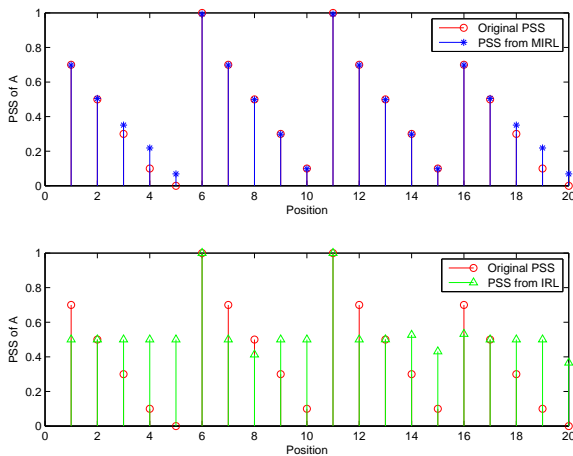


Figure 8. MRL vs IRL on PSS: strong mean and strong covariance

7 CONCLUSION

The experimental results presented in this paper suggest that the MRL problem is worth additional study because learned MRL rewards tend to substantially closer to true rewards and to yield better forward policies than those learned from IRL. Several factors may underlie the performance of MRL. First, a multi-agent system often involves games while IRL assumes the other agents in the environment are passive. Second, from the perspective of game theory, optimal strategies are generally mixed, and these in turn are difficult to handle in IRL. Lastly, IRL cannot fully capture equilibrium information, though some equilibrium information can be reflected in the state transition dynamics.

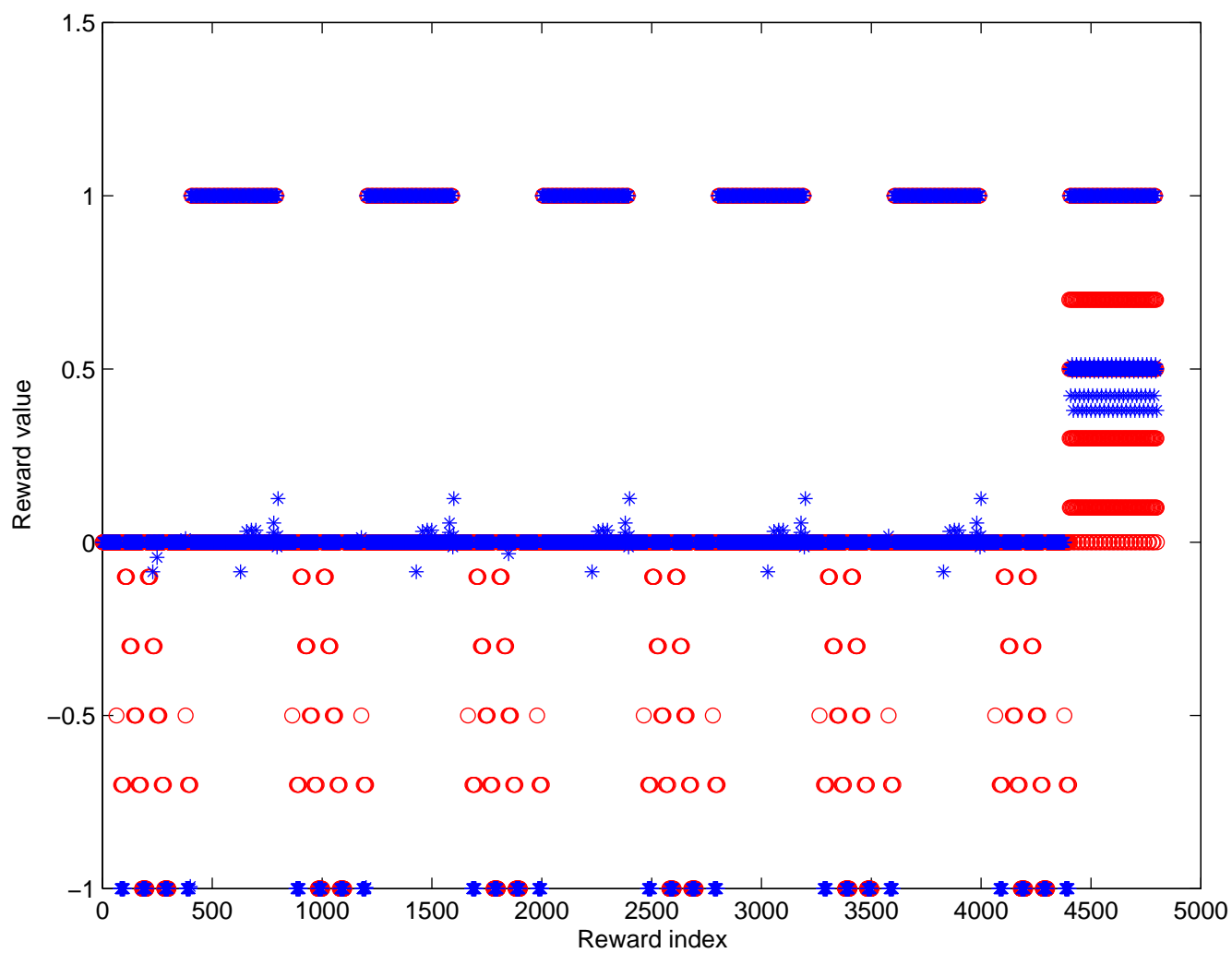
ACKNOWLEDGEMENTS

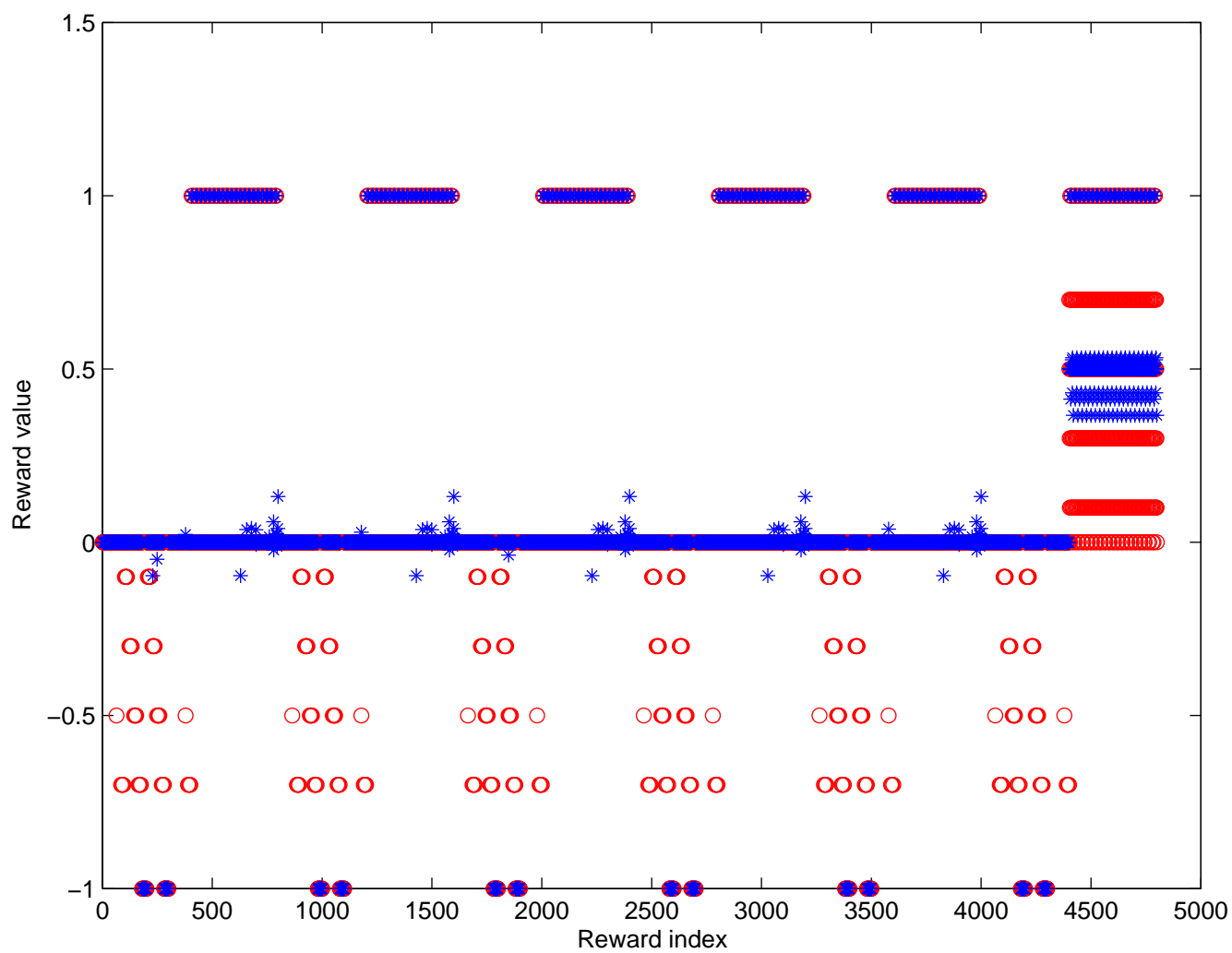
We would like to thank the financial support for this project from Science Applications International Corporation (SAIC) through the Research Scholars Fellowships Program.

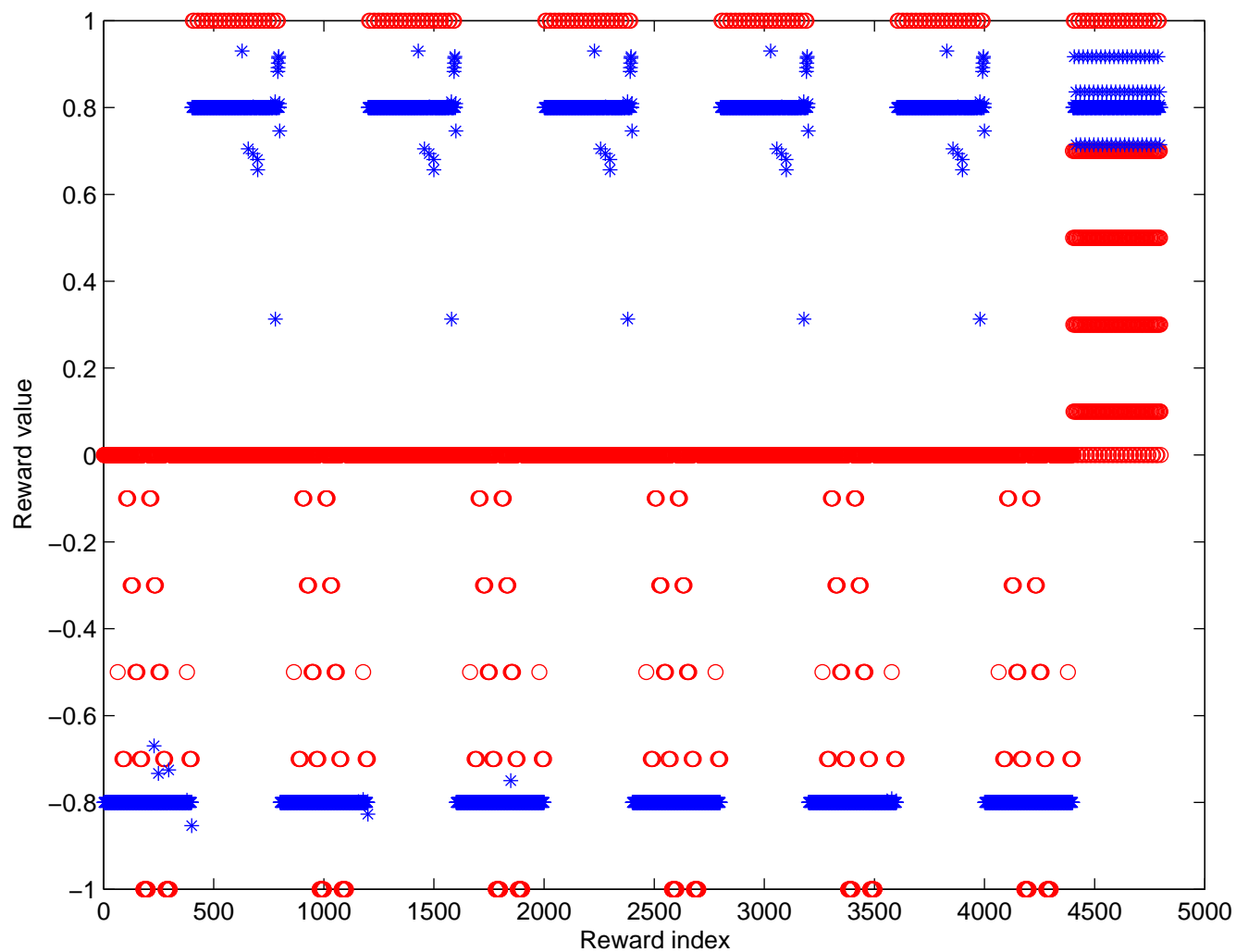
REFERENCES

- [1] S. Abdallah and V. Lesser, ‘A multiagent reinforcement learning algorithm with non-linear dynamics’, *Journal of Artificial Intelligence Research*, **33**, 521–549, (2008).

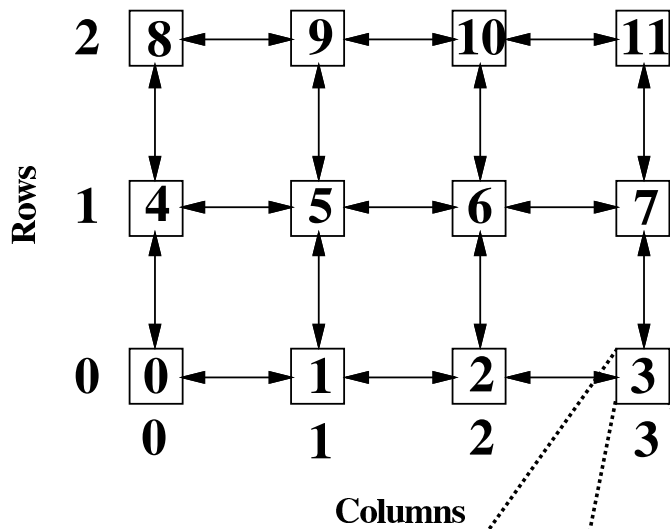
- [2] P. Beling, R. Cogill, and X. Lin., ‘Multiagent inverse reinforcement learning for zero-sum stochastic games’, in *51st Annual Allerton Conference on Communication, Control and Computing*, (2013).
- [3] J. Hu and M. P. Wellman, ‘Multiagent reinforcement learning: Theoretical framework and an algorithm’, in *Proceedings of the 15th International Conference on Machine Learning, ICML’98*, pp. 242–250, (1998).
- [4] I. Kash, E. Friedman, and J. Halpern, ‘Multiagent learning in large anonymous games’, *Journal of Artificial Intelligence Research*, **40**, 571–598, (2011).
- [5] D. Krishnamurthy and E. Todorov, ‘Inverse optimal control with linearly-solvable mdps’, in *Proceedings of the 27th International Conference on Machine Learning, ICML’10*, pp. 335–342, (2010).
- [6] S. Levine, Z. Popović, and V. Koltun, ‘Nonlinear inverse reinforcement learning with gaussian processes’, in *Proceedings of the 24th Advances in Neural Information Processing, NIPS’11*, pp. 19–27, (2011).
- [7] X. Lin, P. A. Beling, and R. Cog, ‘Multi-agent inverse reinforcement learning for zero-sum games’, *CoRR*, (2014). Forthcoming.
- [8] M. L. Littman, ‘Markov games as a framework for multi-agent reinforcement learning’, in *Proceedings of the 11th International Conference on Machine Learning, ICML’94*, pp. 157–163, (1994).
- [9] S. Natarajan, G. Kunapuli, K. Judah, P. Tadepalli, K. Kersting, and J. W. Shavlik, ‘Multi-agent inverse reinforcement learning’, in *Proceedings of the 9th International Conference on Machine Learning and Applications, ICMLA’10*, pp. 395–400, (2010).
- [10] A. Y. Ng and S. Russell, ‘Algorithms for inverse reinforcement learning’, in *Proceedings of the 17th International Conference on Machine Learning, ICML’00*, pp. 663–670, (2000).
- [11] G. Owen, *Game Theory*, W. B. Saunders Company, Philadelphia, PA, 1st edn., 1968.
- [12] S. D. Patek, P. A. Beling, and Y. Zhao, ‘Natural solutions for a class of symmetric games’, in *AAAI Spring Symposium: Game Theoretic and Decision Theoretic Agents*, pp. 47–53, (2007).
- [13] Q. Qiao and P. A. Beling, ‘Inverse reinforcement learning via convex programming’, in *Proceedings of the 2011 American Control Conference, ACC’11*, pp. 113–118, (2011).
- [14] D. Ramachandran and E. Amir, ‘Bayesian inverse reinforcement learning’, in *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI’07*, pp. 2586–2591, (2007).
- [15] K. Waugh, B. Ziebart, and J. Bagnell, ‘Computational rationalization: The inverse equilibrium problem’, in *Proceedings of the 28th International Conference on Machine Learning, ICML’11*, pp. 1169–1176, (2011).
- [16] Y. Zhao, S. Patek, and P. Beling, ‘Decentralized bayesian search using approximate dynamic programming methods’, *IEEE Transactions on Systems, Man and Cybernetics, Part B*, **38**(4), 970–975, (2008).








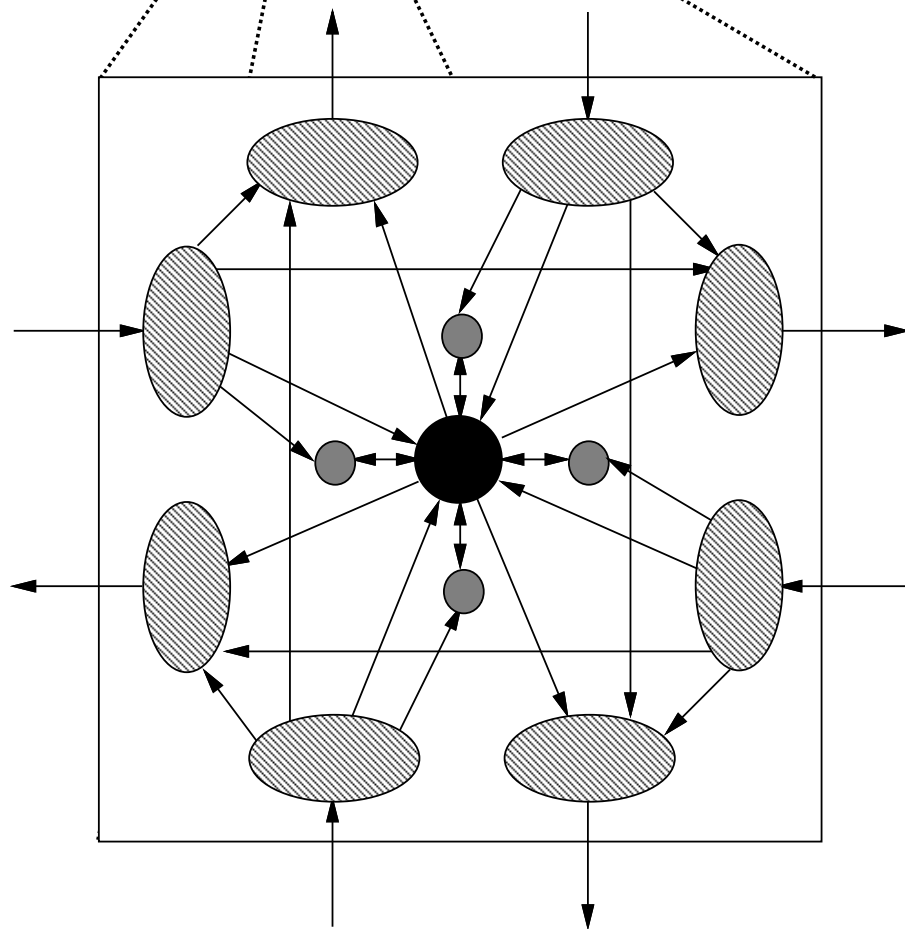


Grid of Transputers



Processes:

-  Routing
-  Buffers
-  Computation



Structure of processes running on each transputer